

International Journal of Population Data Science

Journal Website: www.ijpds.org



An Automated Approach to Develop Crosswalks Between Different Versions of the International Classification of Diseases

Lisa Lix¹, Amani Hamad¹, Yohann Chiu², Mohammad Jafari Jozani¹, Oleguer Plana-Ripoll³, Alessandro Selvitella⁴, and Lin Yan¹

¹University of Manitoba, Winnipeg, Canada

²Universite de Sherbrooke, Sherbrooke, Canada

³Aarhus University, Aarhus, Denmark

⁴Purdue University Fort Wayne, Fort Wayne, USA; eScience Institute - University of Washington, Seattle, USA

Background and Aim

The International Classification of Diseases (ICD) is widely used to record disease diagnoses in administrative data, but different versions of the ICD system are used within and across healthcare systems. Crosswalks for disease coding systems are essential to map equivalent codes between versions. Our aim was to develop and evaluate an automated approach to construct crosswalks for three ICD versions.

Approach

We use ICD-9 clinical modification (ICD-9-CM) as the baseline to construct crosswalks to recent (ICD-10-CA; i.e., Canadian enhancements) and older (ICD-8) versions. Crosswalks were developed for ICD codes in 130 Clinical Classification Software (CCS) adapted categories, which assign chronic condition codes to clinically meaningful groups. We constructed a mapping pipeline using: (1) cosine similarity scores for ICD code labels based on ClinicalBERT, a domain-specific language model, (2) co-occurrence frequency of ICD-9-CM codes with ICD-10-CA and ICD-8 codes in inpatient and outpatient administrative data, and (3) chapter filtering to assess conceptual relatedness. Precision, recall, and accuracy were assessed against manually-constructed crosswalks.

Results and Implications

Manual crosswalks for automated crosswalk evaluation included 331 ICD-9-CM and ICD-8 pairs and 937 ICD-9-CM and ICD-10-CA pairs. Overall performance for the automated ICD-9-CM to ICD-8 crosswalk was: precision 77%, recall 68%, accuracy 56%. Overall performance for the automated ICD-9-CM to ICD-10-CA crosswalk was lower: precision 68%, recall 30%, accuracy 26%. Performance varied across CCS categories, with precision, recall, and accuracy ranging from <20% to 100%. An automated ICD crosswalk approach can minimize mapping burden, but complex crosswalks will benefit from manual review.

